가우스-균일 혼합확률분포의 매개변수 추정에 관한 고찰1)

최성록*, 김태민**, 유원필*
*한국전자통신연구원 로봇연구부
**한국과학기술연구원 전자전산학과
e-mail: sunglok@etri.re.kr

A Study on Parameter Estimation ofGauss-Uniform Probability Distribution

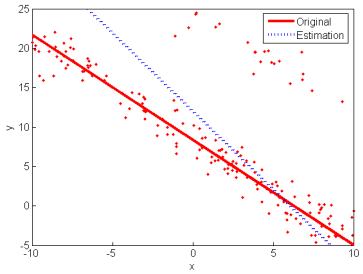
Sunglok Choi*, Taemin Kim**, Wonpil Yu*
*Robot Research Department, ETRI
**Department of Electrical Engineering, KAIST

요 약

본 논문은 완전히 잘못된 데이터가 포함된 회귀(regression) 문제에 사용되는 가우스-균일 혼합확률분포의 두 개의 매개변수 추정에 관하여 고찰한다. 논문에서는 기대값 최대화(Expectation Maximization)와 최우도추정(Maximum Likelihood Estimation)을 이용한 매개변수 추정 방법을 비교한다. 두 기법은 최적화 문제로 기술할 수 있고, 논문에서는 두 기법에서 사용하는 매개변수에 대한 적합도 척도의 개형을 도시하고 비교한다. 몬테-카를로(Monte Carlo) 접근을 통한 두 기법이 추정한 매개변수의 분포를 살펴본다.

1. 서론

공학 분야에서 다뤄지는 많은 문제들 중 하나는 주어진 데이터(data)에서 정보(information)를 얻는 과정이다. 주어진 데이터에서 정해진 모델의 인자들을 추정하는 회귀(regression)는 이러한 문제의 수학적 추상화의 대표적인 예이다. 그러나 실제 많은 경우 주어진 데이터는 그림 1과 같이 완전히 잘못된 데이터, 즉 이상치(outlier)를 포함하고 있고, 이상치는 잘못된 추정을 야기한다. 이상치를 포함한 데이터에서의 강인한 추정 기법에 대한 연구들이 많이 이루어졌는데, 대표적인 것으로 M-estimator, Hough Transform, RANSAC [1] 등이 있다.



(그림 1) 잘못된 데이터가 있는 경우의 최소제곱법

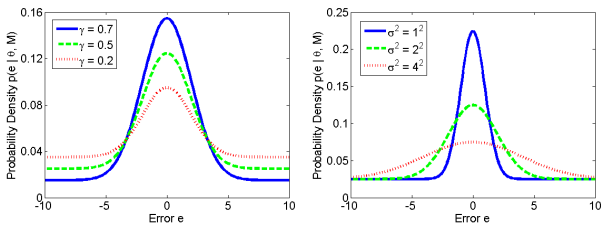
본 논문에서는 에러의 확률분포에 기반을 둔 RANSAC 기법 [2, 3]에 사용되는 가우스-균일 혼합확률분포의 두 매개변수 추정에 대해 고찰한다. 2장에서는 가우스-균일 혼합확률분포에 대해 설명하고, 최우도추정(Maximum Likelihood Estimation, 이하 MLE)과 기대값 최대화(Expectation Maximization, 이하 EM)를 이용한 매개변수 추정에 관하여 기술한다. 3장에서는 두 기법을 통해 추정한 매개변수의 확률분포를 통해 두 기법의 성능을 비교한다.

2. 가우스-균일 혼합확률분포의 매개변수 추정

가우스-균일 혼합확률분포는 작은 노이즈를 가진 올바른 데이터, 즉 정상치(inlier)가 갖는 에러를 가우시안 분포로 가정하고, 이상치가 가지는 에러를 균일 분포로 가정한다. 따라서 가우스-균일 확률분포는 하나의 데이터가 갖는 에러의 확률분포를

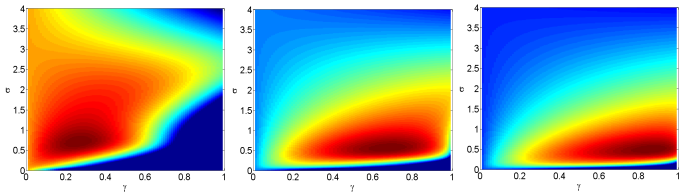
$$p(e|\theta) = \gamma \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{e^2}{2\sigma^2}\right) + (1-\gamma)\frac{1}{\nu} \qquad (1)$$

와 같이 표현한다. ν는 최대로 가능한 에러의 크기이다. 가우스-균일 확률분포는 두 개의 매개변수를 가지고 있다 ($\theta = (\gamma, \sigma)$). 각각 γ는 정상치의 비율을 의미하고, σ는 정상치가 갖는 가우시안 노이즈의 표준편차를 의미한다. 그림 2는 두 매개변수의 변화에 따른 확률분포의 개형을 나타낸 것이다.

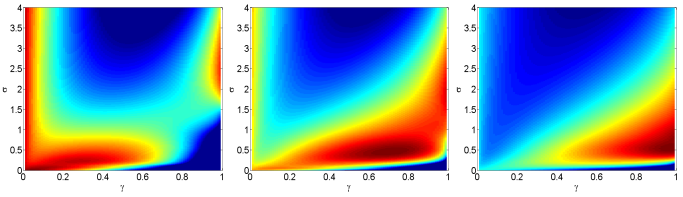


(그림 2) 두 매개변수의 변화에 따른 확률분포의 예 [2]

1) 본 연구는 지식경제부(MKE)와 한국산업기술평가관리원(KEIT)의 연구 과제 (Hybrid u-Robot Service System Technology Development for Ubiquitous City, 2008-S-031-01)의 지원을 받았습니다.



(그림 3) MLE의 $L(\theta)$ ($\gamma=\{0.3,0.6,0.9\}$, $\sigma=0.5$)



(그림 4) EM의 $Q(\theta)$ ($\gamma=\{0.3,0.6,0.9\}$, $\sigma=0.5$)

2.1. 최우도추정(MLE)을 통한 매개변수 추정

MLE는 가장 널리 알려진 매개변수 추정 방법의 하나로 주어진 데이터의 로그우도(likelihood), 즉 주어진 매개변수에 대해 데이터가 있음직한 정도를 최대로 하는 매개변수를 추정한다. 로그우도는 Naive 가정 아래

$$L(\theta)=\sum_{i=1}^{n}\log p(e_i|\theta) \tag{2}$$

와 같고, 개형은 그림 3과 같다. 따라서 MLE를 이용한 매개변수 추정 문제는 $\hat{\theta}=\arg\max_{\theta} L(\theta)$와 같은 최적화 문제가 된다.

2.2. 기대값 최대화(EM)를 통한 매개변수 추정

EM은 관측할 수 없는 변수(latent variable)를 가진 확률분포의 매개변수 추정에 많이 사용되는 기법으로 관측할 수 없는 변수에 대한 로그우도(log likelihood)의 기대값을 최대로 하는 매개변수를 추정한다. 가우스-균일 확률분포에서는 데이터 클래스(데이터가 정상치, 이상치 여부)가 관측할 수 없는 변수가 되고, 식 1을 데이터 클래스 z를 포함한 형태로 표현하면

$$\begin{aligned} p(e|\theta) &= p(e,z=1|\theta)+p(e,z=0|\theta) \\ &= p(z=1|\theta)p(e|z=1,\theta)+p(z=0|\theta)p(e|z=0,\theta) \end{aligned} \tag{3}$$

가 된다. $p(z=1|\theta)$와 $p(z=0|\theta)$는 각각 정상치가 될 사전확률(prior probability)과 이상치가 될 사전확률로 γ, $1-\gamma$이고, $p(e|z=1,\theta)$와$p(e|z=0,\theta)$는 각각 정상치와 이상치에 의한 에러의 확률분포로 식 1의 가우시안 분포와 균일 분포이다. 따라서 로그우도의 기대값은

$$\begin{aligned} Q(\theta;\theta^k) &= \sum_{i=1}^{n} E_{z|e_i,\theta^k}\left(\log p(e_i,z_i|\theta)\right) \\ &= \sum_{i=1}^{n}\left[\sum_{j=0,1} p(z_i=j|e_i,\theta^k)\log p(e_i,z_i=j|\theta)\right] \end{aligned} \tag{4}$$

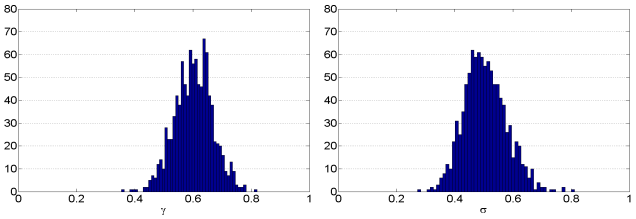
와 같고, 개형은 그림 4과 같다. $p(z_i=j|e_i,\theta^k)$는 i번째 데이터의 클래스가 j일 사후확률(posterior probability)로

$$p(z_i=0|e_i,\theta^k)=\frac{p(z=0|\theta^k)p(e_i|z=0,\theta^k)}{p(e_i|\theta^k)} \tag{5}$$

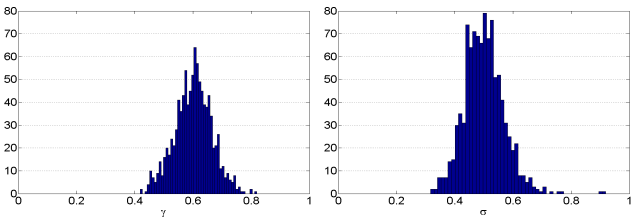
와 같다. 따라서 EM을 이용한 매개변수 추정 문제는 $\hat{\theta}=\arg\max_{\theta} Q(\theta)$와 같은 최적화 문제가 된다.

3. 두 기법의 추정 결과 비교 및 논의

EM의 최대화 과정은 $Q(\theta)$의 미분을 통해 해석적으로 얻을 수 있다 [2]. 그러나 유사한 방법으로 MLE의 최대화 과정을 이루어지지 않는다. 따라서 본 논문에서 MLE의 최대화 과정을 MATLAB의 fmincon 함수를 통해 구현하였다. 그림 5와 6은 임의로 생성된 100개의 데이터에 대해 두 추정 방법을 1000회 실시하여 추정된 두 매개변수의 분포를 나타낸 히스토그램이다. 두 방법 모두 유사한 추정 성능을 나타냈고, 그림 3에서 보는 바와 같은 $Q(\theta)$ 함수의 다봉성(multi-modality) 때문에 EM을 이용한 결과에는 0.8 보다 큰 σ가 2회 추정되었다. 특히 γ가 작아질수록 σ가 커질수록 다봉성은 더욱 뚜렷하게 나타나는데 [2, 3, 4]에서 사용된 EM을 이용한 가우스-균일 확률분포의 매개변수 추정은 이점을 고려하여 수정되어야 할 필요가 있다.



(그림 5) MLE를 통한 추정 결과 (참값: $\gamma=0.6$, $\sigma=0.5$)



(그림 6) EM을 통한 추정 결과 (참값: $\gamma=0.6$, $\sigma=0.5$)

참고문헌

[1] Fischler, M. A. and Rolles, R. C., “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography”, Communications of the ACM, Vol. 24, 1981

[2] Choi, S. and Kim J.-H., “Robust Regression to Varying Data Distribution and Its Application to Landmark-based Localization”, Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 2008

[3] Choi, S., Kim, T., and Yu, W., “Robust Video Stabilization to Outlier Motion using Adaptive RANSAC”, Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2009